14 FLOATING-GATE MOS SYNAPSE TRANSISTORS

Chris Diorio, Paul Hasler, Bradley A. Minch, And Carver Mead

Physics of Computation Laboratory,
California Institute of Technology,
Pasadena, CA 91125, USA

chris@pcmp.caltech.edu

14.1 INTRODUCTION

Our goal is to develop silicon learning systems. One impediment to achieving this goal has been the lack of a simple circuit element combining nonvolatile analog memory storage with locally computed memory updates. Existing circuits [63, 132] typically are large and complex; the nonvolatile floating-gate devices, such as EEPROM transistors, typically are optimized for binary-valued storage [17], and do not compute their own memory updates. Although floating-gate transistors can provide nonvolatile analog storage [1, 15], because writing the memory entails the difficult process of moving electrons through SiO_2, these devices have not seen wide use as memory elements in silicon learning systems.

We have fabricated *synapse transistors* that not only possess nonvolatile analog storage, and compute locally their own memory updates, but also permit simultaneous memory reading and writing, and compute locally the product of their stored memory value and the applied input. To ensure nonvolatile storage, we employ standard floating-gate MOS technology, but we adapt the physical processes that write the memory to perform a local learning function. Although the SiO_2 electron transport still is difficult, and does require high voltages, because our devices integrate both memory storage and local computation within a single device, we expect them to find wide application in silicon learning systems.

We call our devices synapse transistors because, like neural synapses [11], they compute the product of their stored analog memory and the applied input. Also like neural synapses, they can learn from the input signal, without

interrupting the ongoing computation. Although we do not believe that a single device can model the complex behavior of a neural synapse completely, our single-transistor synapses do implement a learning function. With them, we intend to build autonomous learning systems in which both the system outputs, and the memory updates, are computed locally and in parallel.

We have described previously [6, 60, 28] the four-terminal nFET synapse discussed here. We have also described an analog memory cell that employs the nFET device [5], and an autozeroing amplifier that employs the pFET device [12]. We here present the four-terminal nFET synapse in greater detail than we did previously, and for the first time present the four-terminal pFET synapse. We have also described previously a three-terminal nFET synapse [7]. Although the four-terminal synapses require slightly more layout area than does this three-terminal device, the additional terminal gives us greater control over the write and erase processes.

14.2 THE SYNAPSES

The nFET and pFET synapses each possess a poly1 floating gate, a poly2 control gate, and an n-well tunneling implant. Both synapses use hot-electron injection [23] to add electrons to their floating gates, and Fowler-Nordheim (FN) tunneling [16] to remove the electrons. The nFET synapse differs from a conventional n-type MOSFET in its use of a moderately doped channel implant. This implant facilitates hot-electron injection. The pFET synapse, by contrast, achieves a sufficient hot-electron gate current using a conventional p-type MOSFET; no special channel implant is required. We fabricated both synapses in the $2\mu m$ n-well Orbit BiCMOS process available from MOSIS.

In both synapses, the memory is stored as floating-gate charge. Either channel current or channel conductance can be selected as the synapse output. Inputs typically are applied to the poly2 control gate, which couples capacitively to the poly1 floating gate. From the control gate's perspective, altering the floating-gate charge shifts the transistor's threshold voltage V_t, enabling the synapse output to vary despite a fixed-amplitude control-gate input.

We typically operate the synapses in their subthreshold regime [18], and select either drain current or source current as the synapse output. We choose subthreshold operation for three reasons. First, the power consumption of a subthreshold MOSFET typically is less than $1\mu W$. Second, because the channel current in a subthreshold MOSFET is an exponential function of the gate voltage, only small quantities of oxide charge are required for learning. Third, the synapse output is the product of a stored weight and the applied input:

$$I_S = I_o e^{\frac{\kappa V_{fg}}{U_t}} = I_o e^{\frac{\kappa(Q_{fg}+C_{in}V_{in})}{C_T U_t}} = I_o e^{\frac{Q_{fg}}{Q_T}} e^{\frac{\kappa' V_{in}}{U_t}} = W I_o e^{\frac{\kappa' V_{in}}{U_t}} \qquad (14.1)$$

where I_s is the synapse's source current, I_o is the pre-exponential current, κ is the coupling coefficient from the floating gate to the channel, Q_{fg} is the floating-gate charge, C_T is the total capacitance seen by the floating gate, U_t is the

thermal voltage kT/q, C_{in} is the input (poly1 to poly2) coupling capacitance, V_{in} is the control-gate input voltage, $Q_T \equiv C_T U_t/\kappa$, $\kappa' \equiv \kappa C_{in}/C_T$, $W \equiv exp(Q_{fg}/Q_T)$, and, for simplicity, the source potential is assumed to be ground ($V_s = 0$).

The synapse weight W is a learned quantity: Its value derives from the floating-gate charge, which can change with synapse use. The synapse output is the product of W and the source current of an idealized MOSFET that has a control-gate input V_{in} and a coupling coefficient κ' from the control gate to the channel.

Because the tunneling and injection gate currents vary with the synapse terminal voltages and channel current, W varies with the terminal voltages, which are imposed on the device, and with the channel current, which is the synapse output. Consequently, the synapses exhibit a type of learning by which their future output depends on both the applied input and the present output.

14.2.1 The nFET Synapse

Top and side views of the nFET synapse are shown in Fig. 14.1. Its principal features are the following:

- Electrons tunnel from the floating gate to the tunneling implant through the 350Å gate oxide. High voltages applied to the tunneling implant provide the oxide electric field required for tunneling. To prevent breakdown of the reverse-biased pn junction from the substrate to the tunneling implant, we surround the n^+ tunneling implant with a lightly doped n^- well. Tunneling removes electrons from the floating gate, increasing the synapse weight W.

- Electron tunneling is enhanced where the poly1 floating gate overlaps the heavily doped well contact, for two reasons. First, the gate cannot deplete the n^+ contact, whereas it does deplete the n^- well. Thus, the oxide electric field is higher over the n^+. Second, enhancement at the gate edge further augments the oxide field.

- Electrons inject from the drain-to-channel space-charge region to the floating gate. To facilitate injection, we apply a p-type bipolar-transistor base implant to the MOS transistor channel. This implant serves two functions. First, it increases the peak drain-to-channel electric field, thereby increasing the hot-electron population in the drain-to-channel depletion region. Second, it raises the floating-gate voltage, causing the drain-to-gate oxide electric field to favor the transport of injected electrons to the floating gate. Injection adds electrons to the floating gate, decreasing the synapse weight W.

- Oxide uniformity and purity determine the initial matching between synapses, as well as the learning-rate degradations due to oxide trapping. We therefore use the thermally grown gate oxide for all SiO_2 carrier transport.

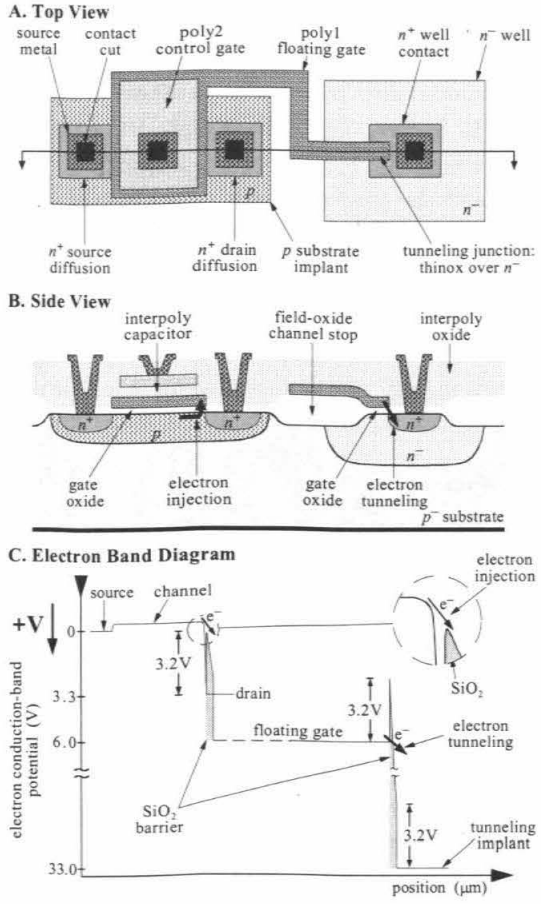

Figure 14.1 The nFET synapse, showing the electron tunneling and injection locations. The three diagrams are aligned vertically. Diagrams A and C are drawn to scale; for clarity, we have exaggerated the vertical scale in diagram B. In the $2\mu m$ Orbit process, the synapse length is $48\mu m$, and the width is $17\mu m$. All voltages in the conduction-band diagram are referenced to the source potential, and we have assumed subthreshold channel currents ($I_s < 100nA$). Although the gate-oxide band diagram actually projects into the plane of the page, for clarity we have rotated it by 90° and have drawn it in the channel direction. When compared with a conventional nFET, the p-type substrate implant quadruples the MOS gate-to-channel capacitance. With a $50fF$ interpoly capacitor as shown, the coupling coefficient between the poly2 control gate and the poly1 floating gate is only 0.2. To facilitate testing, we enlarged the interpoly capacitor to $1pF$, thereby increasing the coupling to 0.8.

14.2.2 The pFET Synapse

Top and side views of the pFET synapse are shown in Fig. 14.2. Its principal features are the following:

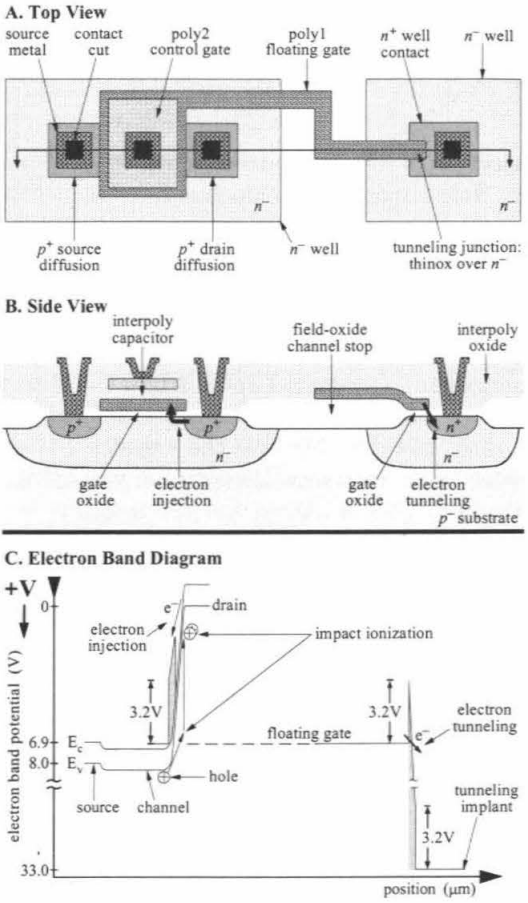

Figure 14.2 The pFET synapse, showing the electron tunneling and injection locations. The well contact is not shown. Like we did in Fig. 14.1, we have aligned the three diagrams vertically, drawn diagrams A and C to scale, exaggerated the vertical scale in diagram B, referenced the voltages in the band diagram to the source potential, and assumed subthreshold ($I_s < 100nA$) operation. Whereas the tunneling process is identical to that in the nFET synapse, the injection process is different. As we describe in the text, we generate the electrons for oxide injection by means of hole impact ionization at the transistor's drain. In the $2\mu m$ Orbit process, the synapse length is $56\mu m$, and the width is $16\mu m$. With a $50fF$ interpoly capacitor as shown, the coupling coefficient between the poly2 control gate and the poly1 floating gate is only 0.25. We enlarged the interpoly capacitor to $1pF$ in the test device, thereby increasing the coupling to 0.8.

- Electrons tunnel from the floating gate to the tunneling implant through the 350Å gate oxide. The tunneling implant is identical to that used in the nFET synapse. As in the nFET synapse, tunneling removes elec-

trons from the floating gate. However, because the pFET and nFET synapses are complementary, tunneling has the opposite effect on the pFET synapse: It decreases, rather than increases, the synapse weight W.

- Electrons inject from the drain-to-channel space-charge region to the floating gate. Hole impact ionization generates the electrons for oxide injection. Channel holes, accelerated in the drain-to-channel electric field, collide with the semiconductor lattice to produce additional electron-hole pairs. The liberated electrons, promoted to their conduction band by the collision, are expelled rapidly from the drain region by this same drain-to-channel electric field. Electrons that acquire more than $3.2eV$ of kinetic energy can, if scattered upward into the gate oxide, inject onto the floating gate. As in the nFET synapse, injection adds electrons to the floating gate; however, because the transistor is a pFET, injection increases, rather than decreases, the synapse weight W.

- Like the nFET synapse, the pFET synapse uses gate oxide for all SiO_2 carrier transport.

14.3 THE GATE-CURRENT EQUATION

We intend to build silicon learning systems using subthreshold synapse transistors. Because the learning behavior of any such system is determined in part by the tunneling and injection processes that alter the stored weights, we have investigated these processes over the subthreshold operating regime.

14.3.1 The Tunneling Process

The tunneling process, for the nFET and pFET synapses, is shown in the energy-band diagrams [9] of Figs. 14.1 and 14.2, respectively. In FN tunneling, a potential difference between the tunneling implant and the floating gate reduces the effective oxide thickness, facilitating electron tunneling from the floating gate, through the SiO_2 barrier, into the oxide conduction band. These electrons are then swept over to the tunneling implant by the oxide electric field. We apply positive high voltages to the tunneling implant to promote electron tunneling.

14.3.2 The Tunneling Equation

The data of Fig. 14.3 show tunneling gate current versus the reciprocal of the voltage across the tunneling oxide. We fit these data with an FN fit [16, 22]:

$$I_g = \varphi V_{ox}^2 e^{-\frac{V_f}{V_{ox}}} \tag{14.2}$$

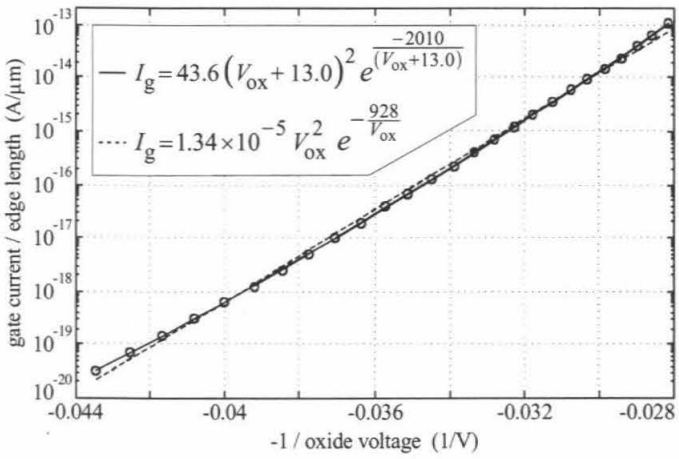

Figure 14.3 Tunneling (gate) current I_g versus $1/V_{ox}$. We define V_{ox} to be the potential difference between the tunneling junction and the floating gate. We normalized the data to the tunneling-junction gate-to-n^+ edge length, in lineal microns, for reasons that we discuss in the text. Our empirical fit (solid line) employs a built-in voltage, V_{bi}, to fit the experimental data more closely; we also show the conventional Fowler-Nordheim fit (dashed line) for comparison.

where I_g is the gate current; V_{ox} is the oxide voltage; $V_f = 928V$ is consistent with a recent survey [18] of SiO_2 tunneling, given the 350Å gate oxide; and φ is a fit parameter. We also show an empirical fit, in which we add a built-in potential, V_{bi}, to the FN equation, to fit the experimental data more closely:

$$I_g = \xi(V_{ox} + V_{bi})^2 e^{-\frac{V_o}{V_{ox}+V_{bi}}} \tag{14.3}$$

where ξ, V_{bi}, and V_o are fit constants.

We normalized the data of Fig. 14.3 to the gate-to-n^+ edge length, in lineal microns, because the floating gate induces a depletion region in the lightly doped n^- well, reducing the effective oxide voltage, and with it the tunneling current. Because the gate cannot appreciably deplete the n^+ well contact, the oxide field is higher where the self-aligned floating gate overlaps the n^+. Because FN tunneling increases exponentially with oxide voltage, tunneling in the synapse transistors is primarily an edge phenomenon.

14.3.3 The Hot-Electron Injection Process

The hot-electron injection process [20], for both the nFET and pFET synapses, is shown in the energy-band diagrams of Figs. 14.1 and 14.2, respectively. Electrons inject from the transistor channel, over the $3.2V$ $Si - SiO_2$ work-function barrier, into the oxide conduction band. These electrons then are swept over

to the floating gate by the oxide electric field. Successful injection, for both the nFET and pFET synapses, requires that the following three conditions be satisfied: (1) the electrons must possess the $3.2eV$ required to surmount the $Si - SiO_2$ barrier, (2) the electrons must scatter upward into the gate oxide, and (3) the oxide electric field must be oriented in the proper direction to transport the electrons to the floating gate.

nFET Injection. In a conventional n-type MOSFET, requirements 1 and 2 are readily satisfied. We merely operate the transistor in its subthreshold regime, with a drain-to-source voltage greater than about $3V$. Because the subthreshold channel-conduction band is flat, the drain-to-channel transition is steep, and the electric field is large. Channel electrons are accelerated rapidly in this field; a fraction of them acquire the $3.2eV$ required for hot-electron injection. A fraction of these $3.2eV$ electrons naturally scatter, by means of collisions with the semiconductor lattice, upward into the gate oxide.

It is principally requirement 3 that prevents injection in a conventional nFET. Subthreshold operation typically implies gate-to-source voltages less than $0.8V$. With the drain at $3V$, and the gate at $0.8V$, the drain-to-gate electric field opposes transport of the injected electrons to the floating gate. The electrons are instead returned to the drain.

In the synapse transistor, we promote the transport of injected electrons to the floating gate by increasing the bulk channel doping. The additional dopant increases the channel surface-acceptor concentration, raising the transistor's threshold voltage from $0.8V$ to $6V$. With the drain at $3V$, and the gate at $6V$, the channel current still is subthreshold, but now the oxide electric field sweeps injected electrons over to the floating gate, rather than returning them to the silicon surface.

pFET Injection. Because the pFET channel current comprises holes, pFET hot-electron injection is different from nFET injection. We accelerate channel holes in the drain-to-channel depletion region of a subthreshold pFET. A fraction of these holes collide with the semiconductor lattice at energies sufficient to liberate additional electron-hole pairs. The ionized electrons, promoted to their conduction band by the collision, are expelled from the drain by the drain-to-channel electric field. If these ionized electrons are expelled with more than $3.2eV$ of kinetic energy, they can inject onto the floating gate.

In the pFET synapse, like in the nFET, injection requirements 1 and 2 are easily satisfied. We merely operate the transistor in its subthreshold regime, with a drain-to-source voltage greater than about $6V$. The higher drain-voltage requirement, when compared with the nFET synapse, is a consequence of the two-step injection process.

In a subthreshold pFET, the gate-to-source voltage typically is less than $1V$; if the drain-to-source voltage exceeds $6V$, the gate voltage must exceed the drain voltage by at least $5V$. The oxide electric field supports strongly the transport of injected electrons to the floating gate, and requirement 3 is always

satisfied. Unlike conventional nFET transistors, conventional pFET transistors naturally inject electrons onto their floating gates (at sufficient drain-to-source voltages); we do not need to add a special channel implant to facilitate injection.

14.3.4 *The Injection Equation*

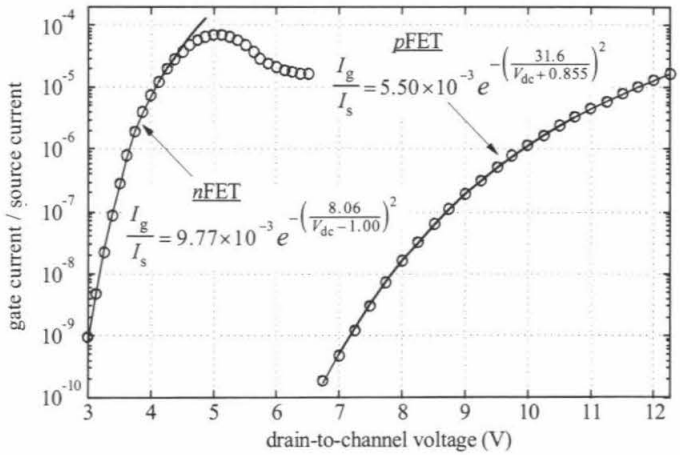

Figure 14.4 Injection efficiency versus drain-to-channel voltage, for both the nFET and pFET synapses. We held the gate-to-channel voltages fixed during the experiments. For the nFET, $V_{gc} = 5.66V$; for the pFET, $V_{gc} = 1.95V$. In the nFET synapse, when the drain voltage exceeds the floating-gate voltage, the oxide E-field tends to return the injected electrons to the silicon surface, rather than transporting them to the floating gate. As a result, for drain-to-channel voltages near $V_{gc} = 5.66V$, the nFET data deviate from the fit.

The data of Fig. 14.4 show injection efficiency (gate current divided by source current) versus drain-to-channel potential, for both the nFET and pFET synapses. The data are plotted as efficiency because, for both devices, the gate current is linearly proportional to the source current over the entire subthreshold range. Because the hot-electron injection probability varies with the drain-to-channel potential, we reference all terminal voltages to the channel. We can re-reference our results to the source terminal using the relationship between source and channel potential in a subthreshold MOSFET [2, 8]:

$$\Psi \approx \kappa V_{fg} + \Psi_o \tag{14.4}$$

where Ψ is the channel potential, V_{fg} is the floating-gate voltage, κ is the coupling coefficient from the floating gate to the channel, and Ψ_0 derives from the MOS process parameters.

For both synapses, the injection efficiency is independent, to first-order, of the floating-gate-to-channel voltage, as long as $V_{fg} > V_d$ (where V_{fg} and V_d

are the floating gate and drain voltages, respectively). In the pFET synapse, this condition is always satisfied. In the nFET synapse, this condition is not necessarily satisfied; the data of Fig. 14.4 show what happens when we sweep the nFET drain from voltages much less than V_{fg}, to voltages much greater than V_{fg}. As V_d approaches V_{fg}, the oxide voltage becomes small, and the gate current drops.

We fit the injection data of Fig. 14.4 empirically; we are currently analyzing the relevant electron-transport physics to derive equivalent analytic results. For the nFET synapse, we chose not to fit the region where $V_d > V_{fg}$ because, at such high drain voltages, the gate currents are too large for use in a practical learning system. For both synapses, then,

$$I_g = \eta I_s e^{-(\frac{V_\beta}{V_{dc}+V_\eta})^2} \tag{14.5}$$

where V_{dc} is the drain-to-channel potential and η, V_β, and V_η are measurable device parameters.

14.3.5 The Gate-Current Equation

Because the tunneling and injection gate currents flow in opposite directions, we obtain the final gate-current equation, for both synapses, by subtracting Eqn. 14.5 from Eqn. 14.2 :

$$I_g = \xi(V_{ox} + V_{bi})^2 e^{-\frac{V_o}{V_{ox}+V_{bi}}} - \eta I_s e^{-(\frac{V_\beta}{V_{dc}+V_\eta})^2} \tag{14.6}$$

The principal difference between the nFET and pFET synapses is the sign of the learning. In the nFET, tunneling increases the weight, whereas injection decreases it; in the pFET, tunneling decreases the weight, whereas injection increases it.

14.3.6 Impact Ionization

We choose source current as the synapse output. Because, for both synapses, the activation energy for impact ionization is less than the barrier energy for injection, a drain-to-channel electric field that generates injection electrons also liberates additional electron-hole pairs [21]. For both synapses, the drain current therefore can exceed the source current. If we choose drain current, rather than source current, as the synapse output, we can rewrite the gate-current equation in terms of drain current using a (modified) lucky-electron [24] formulation:

$$I_d = I_s \left(1 + \varepsilon e^{-\sqrt{\frac{V_m}{V_{ds}-V_\varepsilon}}}\right)$$

where I_d is the drain current and ε, V_m, and V_ε are measurable device parameters. In Fig. 14.5, we plot impact-ionization data for both synapses.

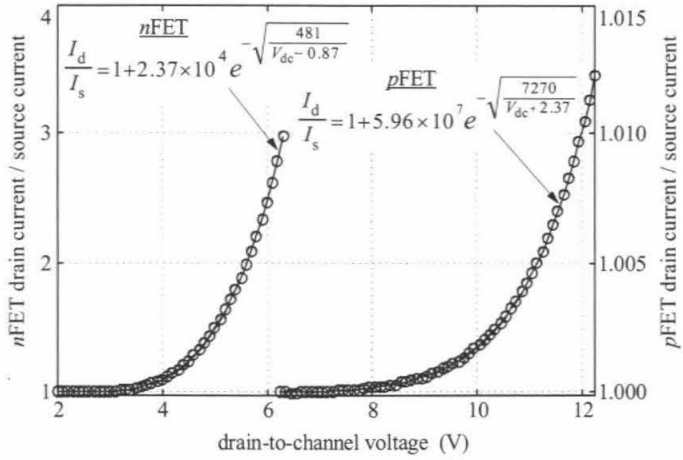

Figure 14.5 Impact ionization versus drain-to-channel potential, for both the nFET and pFET synapses. Impact ionization in the nFET is markedly more efficient than in the pFET, for two reasons. First, as a consequence of its bulk p-type substrate implant, the nFET synapse experiences a higher drain-to-channel electric field than does the pFET, thereby increasing the ionization likelihood. Second, the impact-ionization process is naturally more efficient for electrons (the nFET charge carriers) than it is for holes (the pFET charge carriers).

14.4 SYNAPTIC ARRAYS

A synaptic array, with a synapse transistor at each node, can form the basis of a silicon learning system. We fabricated simplified 2×2 arrays to investigate synapse isolation during tunneling and injection, and to measure the synapse weight-update rates. Because a 2×2 array uses the same row-column addressing employed by larger arrays, it allows us to characterize the synapse isolation and weight-update rules completely.

14.4.1 The nFET Array

The nFET array is shown in Fig. 14.6. We chose, from among the many possible ways of using the array, to select source current as the synapse output, and to turn off the synapses while tunneling. We applied the voltages shown in Table 14.1 to read, tunnel, or inject synapse $\{1, 1\}$ selectively, while ideally leaving the other synapses unchanged.

The tunneling and drain terminals of the array synapse transistors connect within rows, but not within columns. Consequently, the tunneling and injection crosstalk between column synapses is negligible. A synapse's tunneling gate current increases exponentially with the oxide voltage V_{ox}, (V_{ox}, in turn, decreases linearly with V_{fg}), and the hot-electron gate current increases linearly with the channel current I_s, (I_s, in turn, increases exponentially with

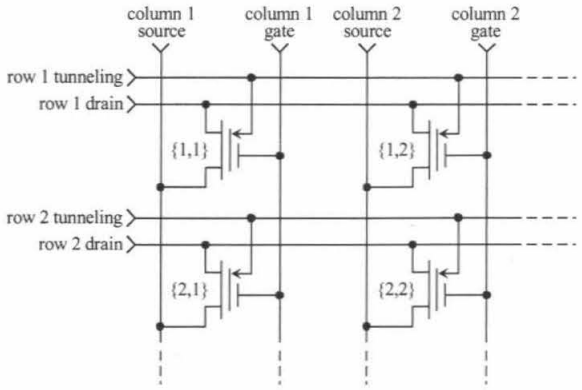

Figure 14.6 A 2×2 array of nFET synapses. Because the row synapses share common tunneling and drain wires, tunneling or injection at one row synapse can cause undesired tunneling or injection at other row synapses.

Table 14.1 The terminal voltages that we applied to the array of Fig. 14.6, to obtain the data of Figs. 14.7 and 14.8.

	col 1 gate	*col 1 source*	*col 2 gate*	*col 2 source*	*row 1 drain*	*row 1 tun*	*row 2 drain*	*row 2 tun*
read	+5	0	0	0	+1	0	0	0
tunnel	0	0	+5	0	0	+31	0	0
inject	+5	0	0	0	3.15	0	0	0

V_{fg}). Consequently, the isolation between row synapses increases exponentially with the voltage differential between their floating gates. By using $5V$ control-gate inputs, we achieve about a $4V$ differential between the floating gates of selected and deselected synapses; the resulting crosstalk between row synapses is $< 0.01\%$ for all operations.

To obtain the data in Fig. 14.7, we initially set all four synapses to $I_s = 100pA$. We tunneled the $\{1,1\}$ synapse up to $100nA$, and then injected it back down to $100pA$, while measuring the source currents of the other three synapses. As expected, the row 2 synapses were unaffected by either the tunneling or the injection. Coupling to the $\{1,2\}$ synapse also was small.

To obtain the data in Fig. 14.8, we first set all four synapses to $I_s = 100nA$. We injected the $\{1,1\}$ synapse down to $100pA$, and then tunneled it back up to $100nA$. As in the experiment of Fig. 14.7, crosstalk to the other synapses was negligible. Our large ($1pF$) gate capacitors provide 80% voltage coupling between a synapse's control and floating gates, minimizing crosstalk at the

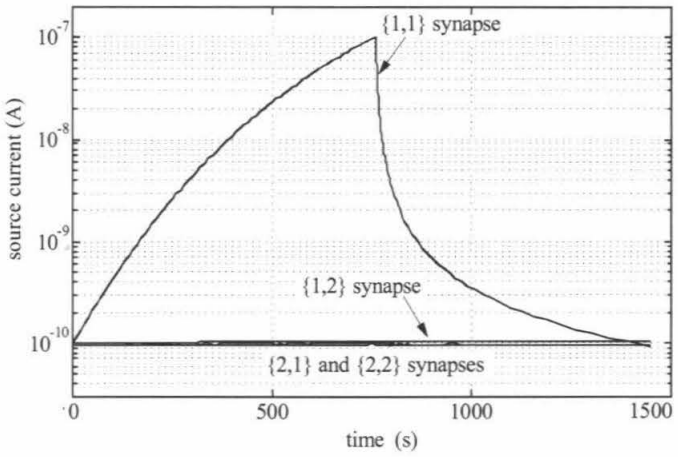

Figure 14.7 Isolation in a 2×2 array of nFET synapses. Source current is the synapse output. The $\{1,1\}$ synapse first is tunneled up to $100nA$, then is injected back down to $100pA$. The tunneling voltage, referenced to the substrate potential, is $V_{tun} = 31V$; the injection voltage is $V_{ds} = 3.15V$. Crosstalk to the $\{1,2\}$ synapse, defined as the fractional change in the $\{1,2\}$ synapse divided by the fractional change in the $\{1,1\}$ synapse, is 0.006% during tunneling, and is 0.002% during injection.

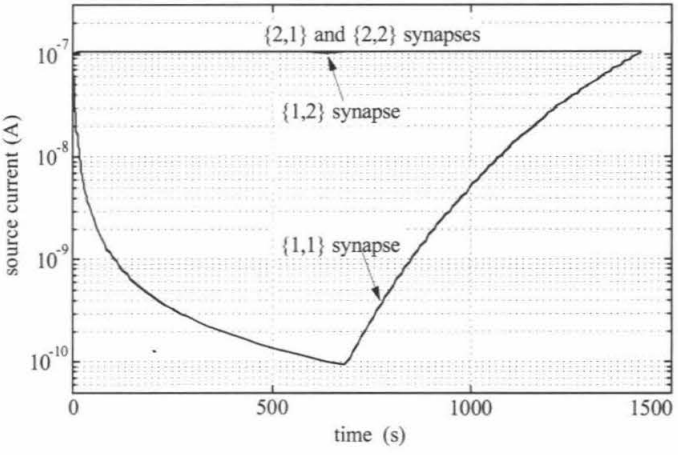

Figure 14.8 Results from the same experiment as in Fig. 14.7, but here the $\{1,1\}$ synapse first is injected down to $100pA$, then is tunneled back up to $100nA$. Crosstalk to the $\{1,2\}$ synapse is 0.001% during injection, and is 0.002% during tunneling.

expense of increased size and decreased weight-update rates. We intend to fabricate future synapses with smaller gate capacitors.

14.4.2 The *pFET Array*

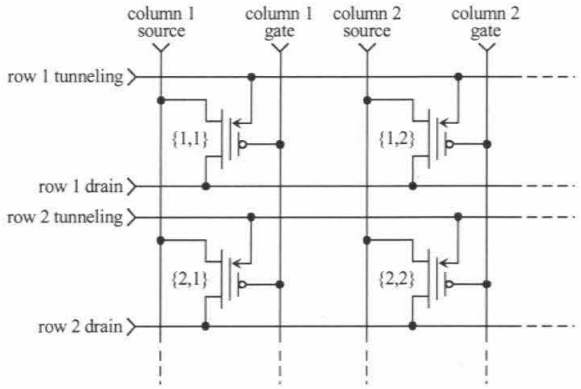

Figure 14.9 A 2×2 array of pFET synapses. The well connections are not shown. As in the nFET array, because the row synapses share common tunneling and drain wires, tunneling or injection at one row synapse can cause undesired tunneling or injection at other row synapses.

The pFET array is shown in Fig. 14.9. We grounded the p-type substrate, applied $+12V$ to the n-type well, and referenced all terminal voltages to the well potential.

Table 14.2 The terminal voltages that we applied to the array of Fig. 14.9, to obtain the data of Figs. 14.10 and 14.11.

	col 1 gate	*col 1 source*	*col 2 gate*	*col 2 source*	*row 1 drain*	*row 1 tun*	*row 2 drain*	*row 2 tun*
read	-5	0	0	0	-5	0	0	0
tunnel	-5	0	0	0	-5	+28	0	0
inject	-5	0	-4	0	-9.3	0	0	0

We again chose source current as the synapse output, but we left the pFET synapses turned on while tunneling, rather than turning them off like we did for the nFET array experiment. We applied the voltages shown in Table 14.2 to read, tunnel, or inject synapse $\{1, 1\}$ selectively, while ideally leaving the other synapses unchanged.

To obtain the data in Fig. 14.10, we initially set all synapses to $I_s = 100pA$. We injected the $\{1, 1\}$ synapse up to $100nA$, and then tunneled it back down to $100pA$. To obtain the data in Fig. 14.11, we performed the opposite experiment.

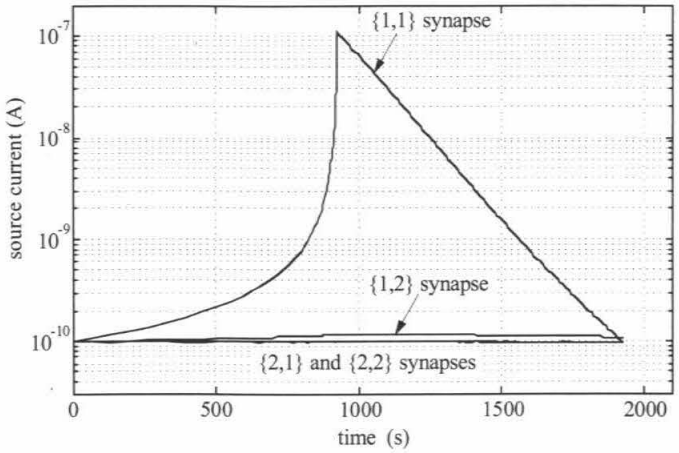

Figure 14.10 Isolation in a 2×2 array of pFET synapses. Source current is the synapse output. The $\{1,1\}$ synapse first is injected up to $100nA$, then is tunneled back down to $100pA$. The injection voltage is $V_{ds} = -9.3V$; the tunneling voltage, referenced to the well potential, is $V_{tun} = 28V$. Crosstalk to the $\{1,2\}$ synapse, defined as the fractional change in the $\{1,2\}$ synapse divided by the fractional change in the $\{1,1\}$ synapse, is 0.016% during injection, and is 0.007% during tunneling.

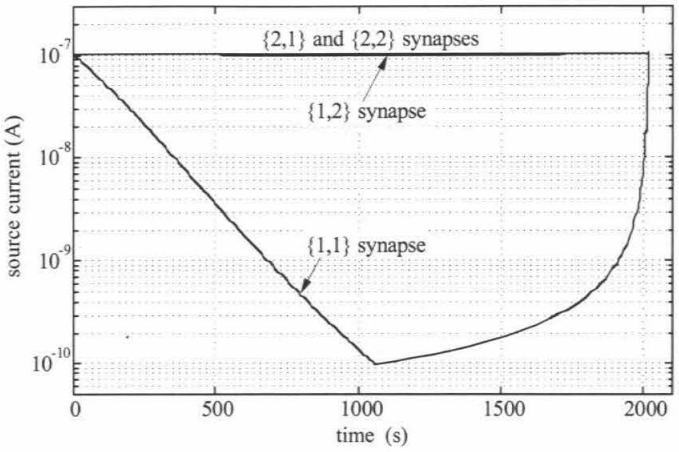

Figure 14.11 Results from the same experiment as in Fig. 14.10, but here the $\{1,1\}$ synapse first is tunneled down to $100pA$, then is injected back up to $100nA$. Crosstalk to the $\{1,2\}$ synapse is 0.005% during injection, and is 0.004% during tunneling.

For the pFET array, like for the nFET array, the crosstalk between column synapses was negligible, and the crosstalk between row synapses was small.

When we injected the $\{1,1\}$ synapse, we applied $-4V$, rather than $0V$, to the $\{1,2\}$ synapse's control gate. We did so because hot-electron injection can occur in a pFET synapse by a mechanism different from that described in Section 14.3: If the floating-gate voltage exceeds the well voltage, and the drain-to-channel potential is large, electrons can inject onto the floating gate by means of a non-destructive avalanche-breakdown phenomenon [23] at the MOS surface.

14.5 THE SYNAPSE WEIGHT-UPDATE RULE

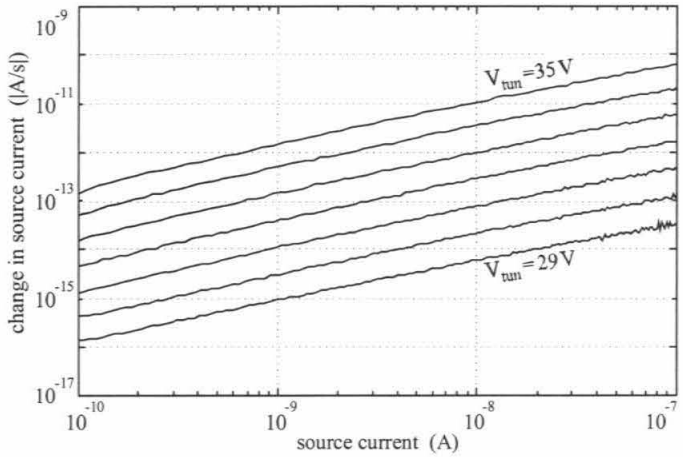

Figure 14.12 The magnitude of the temporal derivative of the source current versus the source current, for an nFET synapse with a continuous tunneling-oxide current. We tunneled the $\{1,1\}$ synapse up as in Fig. 14.7, with the source at ground and the ground-referenced tunneling voltage stepped from $29V$ to $35V$ in $1V$ increments. The mean slope is $+0.83$.

We repeated the experiments of Figs. 14.7 and 14.10, for several tunneling and injection voltages; in Figs. 14.12 through 14.15 we plot, for the nFET and pFET synapses, the magnitude of the temporal derivative of the source current versus the source current. We held the control-gate input V_{in} fixed during these experiments; consequently, the data show the synapse weight updates $\delta W/\delta t$, as can be seen by differentiating Eqn. 14.1. Starting from the gate-current equation, Eqn. 14.6, we now derive weight-update rules that fit these data.

14.5.1 Tunneling

We begin by taking the temporal derivative of the synapse weight W, where $W \equiv exp(Q_{fg}/Q_T)$:

$$\frac{\delta W}{\delta t} = \frac{W}{Q_T}\frac{\delta Q_{fg}}{\delta t} = \frac{W}{Q_T} I_g \tag{14.7}$$

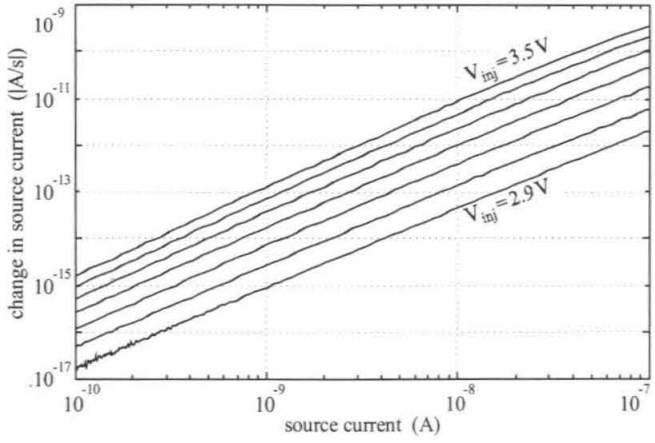

Figure 14.13 The magnitude of the temporal derivative of the source current versus the source current, for an nFET synapse with a continuous hot-electron oxide current. We injected the $\{1, 1\}$ synapse down as in Fig. 14.7, with the source at ground and the ground-referenced drain voltage stepped from $2.9V$ to $3.5V$ in $0.1V$ increments. The mean slope is -1.76; we have added the minus sign because the synapse weight is injecting down.

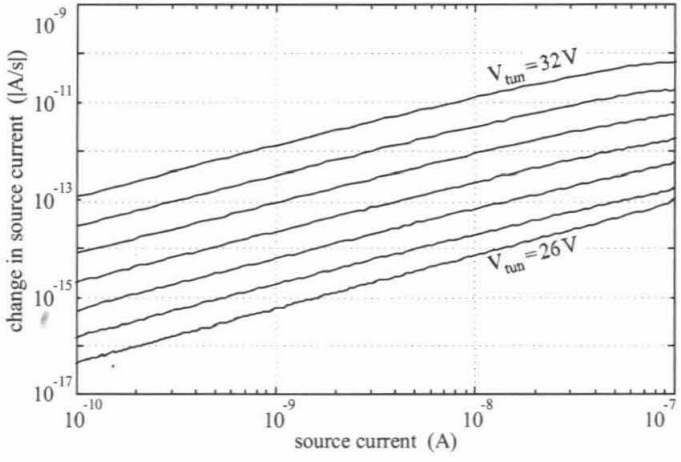

Figure 14.14 The magnitude of the temporal derivative of the source current versus the source current, for a pFET synapse with a continuous tunneling-oxide current. We tunneled the $\{1, 1\}$ synapse down as in Fig. 14.10, with the source and well at $+12V$ and the tunneling voltage, referenced to the well potential, stepped from $26V$ to $32V$ in $1V$ increments. The mean slope is -0.99; we have added the minus sign because the synapse weight is tunneling down.

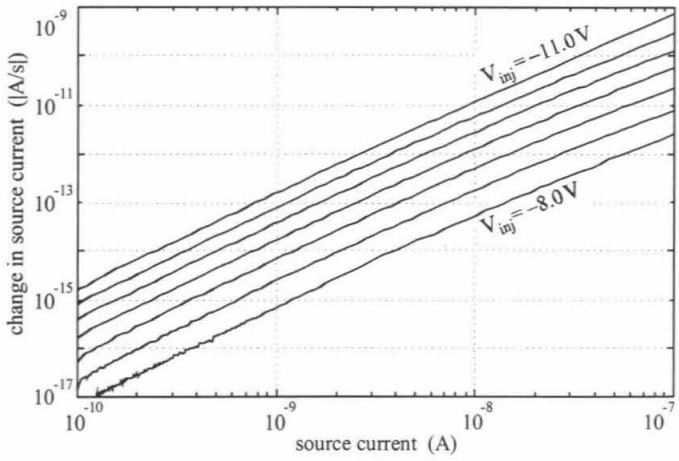

Figure 14.15 The magnitude of the temporal derivative of the source current versus the source current, for a pFET synapse with a continuous hot-electron oxide current. We injected the $\{1, 1\}$ synapse up as in Fig. 14.10, with the source and well at $+12V$ and the drain voltage, referenced to the source potential, stepped from $-8.0V$ to $-11.0V$ in $-0.5V$ increments. The mean slope is $+1.89$.

In Appendix A.1, we substitute for the tunneling gate current using Eqn. 14.3, and solve for the tunneling weight-update rule:

$$\frac{\delta W}{\delta t} \approx \frac{1}{\tau_{tun}} W^{(1-\sigma)} \tag{14.8}$$

where σ and τ_{tun} are defined in Eqns. 14.A.3 and 14.A.4, respectively. Equation 14.8 fits the tunneling weight-update data for both synapses. In the nFET synapse, $0.12 < \sigma < 0.22$; in the pFET, $0.01 < \sigma < 0.05$.

14.5.2 Injection

We begin with $\delta W/\delta t$ from Eqn. 14.7. In Appendix A.2, we substitute for the injection gate current using Eqn. 14.5, and solve for the injection weight-update rule:

$$\frac{\delta W}{\delta t} = -\frac{1}{\tau_{inj}} W^{(2-\varepsilon)} \tag{14.9}$$

where ε and τ_{inj} are defined in Eqns. 14.A.8 and 14.A.9, respectively. Equation 14.9 fits the injection weight-update data for both synapses. In the nFET synapse, $0.14 < \varepsilon < 0.28$; in the pFET, $0.08 < \varepsilon < 0.14$.

14.5.3 The Weight-Update Rule

We obtain the synapse weight-update rule by adding Eqns. 14.8 and 14.9, with a leading $(\pm)$ added because the sign of the updates is different in the nFET and pFET synapses:

$$\frac{\delta W}{\delta t} \approx \pm \left[\frac{1}{\tau_{tun}} W^{(1-\sigma)} - \frac{1}{\tau_{inj}} W^{(2-\varepsilon)} \right] \tag{14.10}$$

For nFET synapses, we use the $(+)$ in Eqn. 14.10; for pFET synapses, we use the $(-)$.

14.5.4 Learning-Rate Degradation

SiO_2 trapping is a well-known issue in floating-gate transistor reliability [3]. In digital EEPROMs, it ultimately limits the transistor life. In the synapses, trapping decreases the weight-update rate. However, because a synapse's weight W is exponential in its floating-gate charge Q_{fg} (see Eqn. 14.1), the synapses in a subthreshold-MOS learning system will transport only small quantities of total oxide charge over the system lifetime. We tunneled and injected $1nC$ of gate charge in both synapses, and measured a $\sim 20\%$ drop in both the tunneling and injection weight-update rates. Because $1nC$ of charge represents an enormous change in synapse weight, we believe that oxide trapping can be ignored safely.

14.6 CONCLUSION

We have described complementary single-transistor silicon synapses with nonvolatile analog memory, simultaneous memory reading and writing, and bidirectional memory updates that are a function of both the applied terminal voltages and the present synapse output. We have fabricated two-dimensional synaptic arrays, and have shown that we can address individual array nodes with good selectivity. We have derived a synapse weight-update rule, and believe that we can build autonomous learning systems, that combine single-transistor analog computation with weight updates computed both locally and in parallel, with these devices. Finally, we anticipate that our single-transistor synapses will allow the development of dense, low-power, silicon learning systems.

Appendix: A

A.1 THE TUNNELING WEIGHT-UPDATE RULE

We begin with the temporal derivative of the synapse weight W (see Eqn. 14.7):

$$\frac{\delta W}{\delta t} = \frac{W}{Q_T} I_g \tag{14.A.1}$$

We substitute Eqn. 14.3 for the gate current I_g:

$$\frac{\delta W}{\delta t} = \frac{\xi W}{Q_T}(V_{ox} + V_{bi})^2 e^{-\frac{V_o}{V_{ox}+V_{bi}}}$$

We substitute $V_{ox} = V_{tun} - V_{fg}$ (where V_{tun} and V_{fg} are the tunneling-node and floating-gate voltages, respectively), assume that $V_{tun} + V_{bi} \gg V_{fg}$, expand the exponent by $(1-x)^{-1} \approx 1 + x$, and solve:

$$\frac{\delta W}{\delta t} \approx \frac{\xi W}{Q_T}(V_{tun} + V_{bi} - V_{fg})^2 e^{\frac{-V_o}{V_{tun}+V_{bi}} - \frac{V_o V_{fg}}{(V_{tun}+V_{bi})^2}}$$

We substitute $V_{fg} = U_t Q_{fg}/\kappa Q_T$, and solve for the tunneling weight-update rule:

$$\frac{\delta W}{\delta t} \approx \frac{\xi}{Q_T}(V_{tun} + V_{bi} - V_{fg})^2 e^{\frac{-V_o}{V_{tun}+V_{bi}}} W^{(1-\sigma)} \tag{14.A.2}$$

where

$$\sigma \equiv \frac{V_o U_t}{\kappa (V_{tun} + V_{bi})^2} \tag{14.A.3}$$

Because, for subthreshold source currents, the floating-gate voltage changes slowly, we approximate $(V_{tun} + V_{bi} - V_{fg})^2$ to be a constant, independent of W, and define

$$\tau_{tun} \equiv \frac{Q_T}{\xi}(V_{tun} + V_{bi} - V_{fg})^{-2} e^{\frac{V_o}{V_{tun}+V_{bi}}} \tag{14.A.4}$$

Finally, we substitute τ_{tun} into Eqn. 14.A.2, to get the tunneling weight-update rule:

$$\frac{\delta W}{\delta t} \approx \frac{1}{\tau_{tun}} W^{(1-\sigma)}$$

A.2 THE INJECTION WEIGHT-UPDATE RULE

We begin by rewriting a synapse transistor's drain-to-channel potential, V_{dc}, in terms of V_{ds} and I_s. In a subthreshold floating-gate MOSFET, the source current is related to the floating-gate and source voltages [18] by

$$I_s = I_o e^{\frac{\kappa V_{fg} - V_s}{U_t}} \tag{14.A.5}$$

Using Eqns. 14.4 and 14.A.5, we solve for the surface potential Ψ in terms of I_s and V_s:

$$\Psi = V_s + \Psi_o + U_t \ln\left(\frac{I_s}{I_o}\right)$$

We now solve for V_{dc}:

$$V_{dc} = V_d - \Psi = V_{ds} - \Psi_0 - U_t \ln\left(\frac{I_s}{I_o}\right) \tag{14.A.6}$$

The injection gate current I_g is given by Eqn. 14.5. We add a minus sign to the gate current, because hot-electron injection decreases the floating-gate charge, and substitute for V_{dc} using Eqn. 14.A.6:

$$I_g = -\eta I_s e^{-\left(\frac{V_\beta}{V_{ds}+V_\eta-\Psi_o-U_t \ln\left(\frac{I_s}{I_o}\right)}\right)^2}$$
$$= -\eta I_s e^{-\left(\frac{V_\beta}{V_{ds}+V_\eta-\Psi_o}\right)^2\left[1-\frac{U_t}{V_{ds}+V_\eta-\Psi_o}\ln\left(\frac{I_s}{I_o}\right)\right]^{-2}}$$

We expand the exponent by $(1-x)^{-2} \approx 1 + 2x$, substitute for I_s using Eqn. 14.1, and solve:

$$I_g \approx -\eta I_o e^{\left[\frac{(1-\varepsilon)\kappa' V_{in}}{U_t} - \left(\frac{V_\beta}{V_{ds}+V_\eta-\Psi_o}\right)^2\right]} W^{(1-\varepsilon)} \tag{14.A.7}$$

where

$$\varepsilon \equiv \frac{2U_t V_\beta^2}{(V_{ds}+V_\eta-\Psi_o)^3} \tag{14.A.8}$$

We substitute Eqn. 14.A.7 into $\delta W/\delta t$, Eqn. 14.A.1,

$$\frac{\delta W}{\delta t} = -\frac{\eta I_o}{Q_T} e^{\left[\frac{(1-\varepsilon)\kappa' V_{in}}{U_t} - \left(\frac{V_\beta}{V_{ds}+V_\eta-\Psi_o}\right)^2\right]} W^{(2-\varepsilon)}$$

We define

$$\tau_{inj} \equiv \frac{Q_T}{\eta I_o} e^{\left[\left(\frac{V_\beta}{V_{ds}+V_\eta-\Psi_o}\right)^2 - \frac{(1-\varepsilon)\kappa' V_{in}}{U_t}\right]} \tag{14.A.9}$$

Finally, we substitute τ_{inj} into Eqn. 14.A.9 to get the injection weight-update rule:

$$\frac{\delta W}{\delta t} = -\frac{1}{\tau_{inj}} W^{(2-\varepsilon)}$$

Acknowledgments

Lyn Dupré edited the manuscript. This work was supported by the Office of Naval Research, by the Advanced Research Projects Agency, by the Beckman Hearing Institute, by the Center for Neuromorphic Systems Engineering as a part of the National Science Foundation Engineering Research Center Program, and by the California Trade and Commerce Agency, Office of Strategic Technology.

References

[1] T. Allen, A. Greenblatt, C. Mead, and J. Anderson. Writeable analog reference voltage storage device. *U.S. Patent No. 5,166,562*, 1991.

[2] A. G. Andreou and K. A. Boahen. Analog VLSI signal and information processing. In M. Ismail and T. Fiez, editors, *Neural information processing II*, pages 358–413. McGraw-Hill, New York, 1994.

[3] S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masuoka. Reliability issues of flash memory cells. In *Proc. of the IEEE*, volume 82-5, pages 776–787, 1993.

[4] P. Churchland and T. Sejnowski. *The Computational Brain.* MIT Press, 1993.

[5] C. Diorio, P. Hasler, B. A. Minch, and C. Mead. A high-resolution nonvolatile analog memory cell. In *Proc. IEEE Intl. Symp. on Circuits and Systems*, volume 3, pages 2233–2236, 1995.

[6] C. Diorio, P. Hasler, B. A. Minch, and C. Mead. A single transistor silicon MOS device for long term learning. *U.S. Patent Office serial no. 08/399966*, Mar. 7 1995.

[7] C. Diorio, P. Hasler, B. A. Minch, and C. Mead. A single-transistor silicon synapse. *IEEE Trans. Electron Devices*, 43(11):1972–1980, 1996.

[8] C. C. Enz, F. Krummenacher, and E. A. Vittoz. An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications. *Analog Integrated Circuits and Signal Processing*, 8:83–114, 1995.

[9] A. S. Grove. *Physics and Technology of Semiconductor Devices.* John Wiley & Sons, New York, 1967.

[10] P. Hasler, C. Diorio, B. A. Minch, and C. Mead. Single transistor learning synapses. In *Advances in Neural Information Processing Systems 7*, pages 817–824. MIT Press, Cambridge, MA, 1995.

[11] P. Hasler, C. Diorio, B. A. Minch, and C. Mead. Single transistor learning synapses with long term storage. In *IEEE Intl. Symp. on Circuits and Systems*, volume 3, pages 1660–1663, 1995.

[12] P. Hasler, B. A. Minch, C. Diorio, and C. Mead. An autozeroing amplifier using pfet hot-electron injection. In *Proc. IEEE Intl. Symp. on Circuits and Systems*, volume 3, pages 325–328, Atlanta, May 1996.

[13] B. Hochet, V. Peiris, S. Abdo, and M. J. Declercq. Implementation of a learning kohonen neuron based on a new multilevel storage technique. *IEEE J. Solid-State Circuits*, 26(3):262–267, 1991.

[14] P. Hollis and J. Paulos. A neural network learning algorithm tailored for VLSI implementation. *IEEE Tran. Neural Networks*, 5(5):784–791, 1994.

[15] J. Lazzaro, J. Wawrzynek, , and A. Kramer. Systems technologies for silicon auditory models. *IEEE Micro*, 14(3):7–15, June 1994.

[16] M. Lenzlinger and E. H. Snow. Fowler-nordheim tunneling into thermally grown sio_2. *J. of Appl. Phys.*, 40(6):278–283, 1996.

[17] F. Masuoka, R. Shirota, and K. Sakui. Reviews and prospects of non-volatile semiconductor memories. *IEICE Trans.*, E 74(4):868–874, 1991.

[18] C. Mead. Scaling of MOS technology to submicrometer feature sizes. *J. of VLSI Signal Processing*, 8:9–25, 1994.

[19] C. A. Mead. *Analog VLSI and Neural Systems.* Addison-Wesley, Reading, MA, 1989.

[20] J. J. Sanchez and T. A. DeMassa. Review of carrier injection in the silicon/silicon-dioxide system. In *IEE Proceedings-G*, volume 138-3, pages 377–389, 1991.

[21] W. Shockley. Problems related to pn junctions in silicon. *Solid-State Electronics*, 2(1):35–67, 1961.

[22] S. M. Sze. *Physics of Semiconductor Devices.* John Wiley & Sons, New York, 1981.

[23] E. Takeda, C. Yang, and A. Miura-Hamada. *Hot-Carrier Effects in MOS Devices.* Academic Press, San Diego, CA, 1995.

[24] S. Tam, P. Ko, and C. Hu. Lucky-electron model of channel hot-electron injection in MOSFET's. *IEEE Trans. Electron Devices*, ED-31(9):1116–1125, 1984.